BMC Genomics

BioMed Central

Research article

Open Access

Novel conserved domains in proteins with predicted roles in eukaryotic cell-cycle regulation, decapping and RNA stability

Vivek Anantharaman and L Aravind*

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Email: Vivek Anantharaman - ananthar@mail.nih.gov; L Aravind* - aravind@mail.nih.gov

* Corresponding author

Published: 16 July 2004

BMC Genomics 2004, **5**:45 doi:10.1186/1471-2164-5-45

Received: 27 February 2004

Accepted: 16 July 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/45

Abstract

Background: The emergence of eukaryotes was characterized by the expansion and diversification of several ancient RNA-binding domains and the apparent *de novo* innovation of new RNA-binding domains. The identification of these RNA-binding domains may throw light on the emergence of eukaryote-specific systems of RNA metabolism.

Results: Using sensitive sequence profile searches, homology-based fold recognition and sequence-structure superpositions, we identified novel, divergent versions of the Sm domain in the Scd6p family of proteins. This family of Sm-related domains shares certain features of conventional Sm domains, which are required for binding RNA, in addition to possessing some unique conserved features. We also show that these proteins contain a second previously uncharacterized C-terminal domain, termed the FDF domain (after a conserved sequence motif in this domain). The FDF domain is also found in the fungal Dcp3p-like and the animal FLJ22128-like proteins, where it fused to a C-terminal domain of the YjeF-N domain family. In addition to the FDF domains, the FLJ22128-like proteins contain yet another divergent version of the Sm domain at their extreme N-terminus. We show that the YjeF-N domains represent a novel version of the Rossmann fold that has acquired a set of catalytic residues and structural features that distinguish them from the conventional dehydrogenases.

Conclusions: Several lines of contextual information suggest that the Scd6p family and the Dcp3p-like proteins are conserved components of the eukaryotic RNA metabolism system. We propose that the novel domains reported here, namely the divergent versions of the Sm domain and the FDF domain may mediate specific RNA-protein and protein-protein interactions in cytoplasmic ribonucleoprotein complexes. More specifically, the protein complexes containing Sm-like domains of the Scd6p family are predicted to regulate the stability of mRNA encoding proteins involved in cell cycle progression and vesicular assembly. The Dcp3p and FLJ22128 proteins may localize to the cytoplasmic processing bodies and possibly catalyze a specific processing step in the decapping pathway. The explosive diversification of Sm domains appears to have played a role in the emergence of several uniquely eukaryotic ribonucleoprotein complexes, including those involved in decapping and mRNA stability.

Background

Systematic comparative analyses of genome sequences have suggested that the majority of domains found in proteins involved in RNA metabolism are drawn from a relatively small set of conserved domains (approximately 100–135) [1-3]. The proteins containing these conserved domains correspond to around 4 to 11 percent of the protein-coding genes in cellular life forms and perform a wide range of functions that include translation and its regulation, processing and modification of cellular RNAs, and post-transcriptional gene regulation [1-3]. This set of conserved domains can be broadly divided into those that mediate interactions with RNAs or other proteins in ribonucleoprotein complexes, and catalytic domains that may catalyze a wide range of reactions related to RNA or associated proteins. Most of the common RNA-binding domains (RBDs) are relatively small (less than 150 residues) and tend to be evolutionarily mobile, occurring as solos, or in combination with other RBDs or enzymatic domains [1]. Several RBDs as well as the catalytic domains of RNA metabolism enzymes are amongst the most highly conserved and universally distributed protein domains in cellular organisms. These highly conserved domains are typically present in ribosomal components, translation factors, enzymes that modify rRNA and tRNA, polyadenylation, and transcription elongation factors [1,4,5]. However, the analysis of phyletic patterns of conserved domains has also suggested that a significant innovation of novel RBDs occurred at the base of eukaryotes [1]. These eukaryotic innovations include the PAZ, G-Patch, PWI and SWAP domains and several Zn-chelating domains, such as the Zn-knuckle, the CCCH and LRP fingers [1,6-8]. The emergence of these domains, as well as the expansion and diversification of superfamilies of previously existing domains appears to have accompanied development of several novel aspects of RNA metabolism in the eukaryotes. These unique eukaryotic aspects include pathways involved in pre-mRNA splicing, capping, post-transcriptional gene silencing and nucleo-cytoplasmic RNA transport. The eukaryotes also possess more complex versions of RNA degradation and processing systems, such as the exosome and the multi-subunit RNaseP/RNase MRP [1,9,10]. Hence, the identification of novel eukaryote-specific domains, as well as the analysis of the diversification of ancient domain superfamilies in eukaryotes may help in providing a better understanding of the origins and the biochemical properties of the unique aspects their RNA metabolism.

The computational identification of conserved RNA-binding domains (RBDs) has considerably contributed to the analysis of RNA-protein interactions in various pathways of RNA metabolism [1,6,11-13]. The enzymatic domains associated with RNA metabolism typically belong to superfamilies, which may also include members that act on substrates outside the context of RNA metabolism (eg. Rossmann fold methyltransferases acting on non-ribonucleoprotein substrates) [1]. Hence, the combinations of RBDs and enzymatic domains in the same polypeptide provide a strong contextual handle for predicting novel catalytic activities associated with RNA metabolism. Comprehensive analysis of the commonly occurring domains involved in RNA metabolism has previously helped in identifying several such domain architectures that led to the prediction of novel RNA and RNP modifying/processing enzymes [1,6,14,15]. The recent increase in the available genomic sequences from eukaryotes provides further opportunities to extract contextual information in the form of previously unnoticed domain architectures. Furthermore, the new data also allows the detection of less common, nevertheless functionally important eukaryote-specific domains, which may have eluded earlier screens for such domains. Additionally, other forms of contextual information emerging from newer studies involving large-scale mutational analysis of eukaryotic genes, high-throughput analysis of gene expression, sub-cellular protein localization and protein-protein interactions could also provide clues regarding the functions of uncharacterized proteins.

In particular, we are interested in using computational methods to identify novel eukaryote-specific proteins that may be involved in RNA metabolism and predicting their potential biochemical functions. In the current work we use a combination of sequence analysis, homology-based fold prediction and contextual information to describe two novel conserved RNA-protein or protein-protein interaction modules and one catalytic module that are found in proteins predicted to participate in regulation of the cell cycle and decapping. We discuss these findings in the context of the origin of the decapping apparatus in eukaryotes and present hypotheses for the possible functions of poorly characterized but highly conserved groups of eukaryotic proteins.

Results and discussion

Identification of the novel FDF domain and conserved eukaryotic proteins with domains related to the RNA-binding domain SM domain

Several RNA binding proteins in eukaryotes are characterized by the presence of highly charged or polar low-complexity segments, typically containing repeats of simple motifs such as SR, RG and GGY [16-18]. Experimental evidence has suggested that these segments interact with RNA with low target specificity or aid in their localization to specific RNA processing substructures [16-19]. These segments are usually combined with globular domains that may mediate more specific interactions with RNA. Hence, detection of proteins containing these segments provides a means of identifying potential RNA-binding

proteins that may either lack previously characterized RBDs or contain very divergent versions of them. Accordingly, we generated a sieve for such proteins using pattern searches that identified proteins with multiple occurrences of the low-entropy repeat motifs that are typical of RNA-binding proteins. Those proteins in this set, which were identified as potential RNA-binding proteins or RNA-processing enzymes in our previous surveys [1,20,21] conducted using sensitive profiles for RBDs and associated enzymes, were removed in the first step. Of the proteins that remained, we selected those proteins that contained potential globular domains when screened using the SEG program [22]. These proteins were then further searched using the PFAM domain collection [23] to identify any previously reported modules that may have escaped our searches.

Via this procedure we identified one group of experimentally uncharacterized proteins typified by *Saccharomyces cerevisiae* Scd6p and *Schizosaccharomyces pombe* Sum2p as potential RNA-binding proteins. These proteins formed a distinctive family (hereinafter Scd6p family), which included the mRNA binding protein Rap55 from the newt *Pleurodeles waltl* and orthologous representatives from fungi, animals, plants and apicomplexans (*Cryptosporidium* and *Plasmodium*). This observation suggests that the family is likely to have emerged prior to the diversification of the crown group eukaryotes and possibly performs a well-conserved function. Analysis with the SEG program [22] suggested that these proteins contain distinct N- and C-terminal globular domains flanked by low complexity regions enriched in charged residues, including the RS and RG motifs. In order to understand better the affinities of these globular domains we initiated PSI-BLAST searches (profile inclusion threshold = .01; iterated to convergence) of the Non-Redundant database (NR) with them using representatives from several different organisms. Interestingly, in searches with the N-terminal module, Sm RNA-binding domains were recovered, either with significant hits ($e = 10^{-4}$–10^{-6}) or as the best hits with borderline E-values. As these domains had not been reported by others or us in systematic surveys for Sm proteins [1,24], we investigated them in greater detail using new position-specific score matrices, which were made by including all the previously identified representatives of Sm domains in the nr database. A search of the NR database with this profile recovered members of the Scd6p in iteration 7 with significant e-values ($e = 10^{-4}$–10^{-6} at the point of first recovery). Secondary structure prediction using a multiple alignment of the of the N-terminal globular domain of the Scd6p family showed that it possessed an all β-fold with a perfect correspondence to the secondary structure elements observed in the Sm-type SH3 β-barrel fold [25,26] (also see SCOP Database [27]). Barring the Sm domains, neither other members of the SH3-like folds nor any other distinct β-barrel-folds, such as the OB fold, were recovered in these searches. Likewise, the Scd6p-like proteins were not detected in searches with profiles for various OB fold domains and other β-strand rich RNA-binding domains. These observations strongly suggested that the Scd6p family contained a previously unreported, divergent form of the Sm domain.

A multiple alignment of the classical Sm domain was generated using a structural superposition of all crystallized Sm domains proteins from the PDB database, including the divergent bacterial version Hfq, as a template (Fig. 1). A comparison of the multiple alignment of the Scd6p family with this alignment of Sm domains shows that it contains the hall mark features of the latter class, such as the presence of a hxG signature (where h is a hydrophobic residue) in the N-terminal half and a +Gpph signature (where 'p' is a polar residue and '+' a positively charged residue), which is seen in the C-terminal half of the archaeo-eukaryotic versions (Fig. 1). Additionally the Scd6p family contains certain unique features that set it apart from other Sm domains: 1) It contains a conserved C-terminal extension that is likely to form an additional terminal strand that is usually lacking in many of the classical Sm domains (Fig. 1). 2) It contains a characteristic motif, usually of the form GTEx+ (where + is a positively charged residue; x is any residue) in the variable region separating the conserved N- and C-terminal halves of the Sm domain (Fig. 1). Most Sm domains contain a helix of variable length at their N-terminus [28]. The Scd6p family shows relatively poor sequence conservation and weak helix prediction in the corresponding N-terminal regions. However, the presence of the conservation in the Scd6p family of the capping residue (either glycine or a small residue), which is present in the C-terminus of this helix, suggests that it might contain an abbreviated version of this helix (Fig. 1).

The Sm proteins from archaea and eukaryotes and the bacterial Hfq proteins do not bind RNAs stably as monomers, but only as heptameric or hexameric toroids [25,28]. Furthermore, even the highly divergent versions of the Sm superfamily, such as the MscS protein of the bacterial mechano-sensory channels [29], form heptameric toroids similar to the RNA binding Sm domains, suggesting that this quaternary structure may be pervasive throughout this superfamily. Accordingly, we speculate that the Scd6p proteins are also likely to be incorporated into such structures. When the conservation pattern of the Scd6p proteins is compared to the RNA contacts of the Sm domains in the crystal structure of the *Archaeoglobus fulgidus* Sm1 (AF0875) heptameric ring, several similarities and a few notable differences are seen [25] (Fig. 1). In the highly conserved C-terminal +Gpph motif, the side chain of the positively charged residue (R63 in Af Sm1/AF0875)

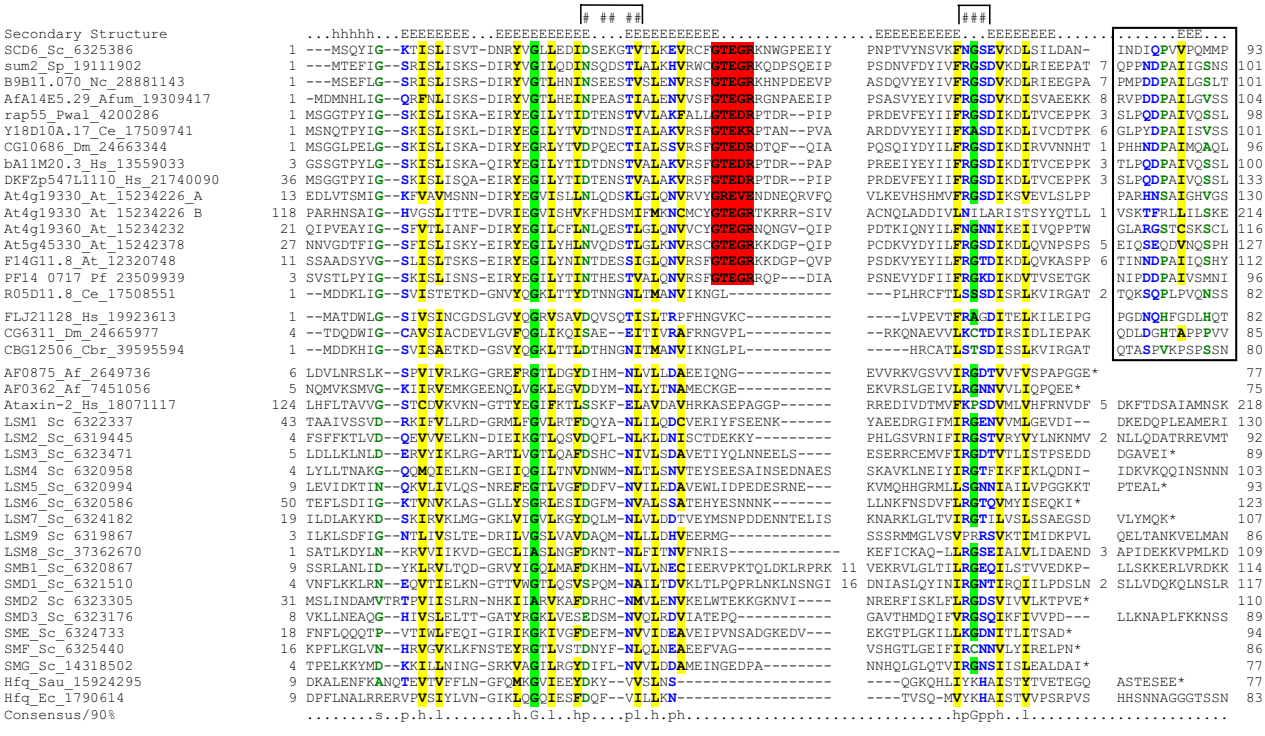

Figure 1

Multiple alignment of the Scd6p family with representatives of other Sm domains. Multiple sequence alignment of the Sm domain of the Scd6p family was constructed using T-Coffee after parsing high-scoring pairs from PSI-BLAST search results. The secondary structure from the crystal structures is shown above the alignment with E representing a strand. The 90% consensus shown below the alignment was derived using the following amino acid classes: hydrophobic (h: ALICVMYFW, yellow shading) and its aliphatic subset (l: ALIV, yellow shading); small (s: ACDGNPSTV, green); and polar (p: CDEHKNQRST, blue). The limits of the domains are indicated by the residue positions, on each end of the sequence. A '*' denotes the end of the protein sequence. The numbers within the alignment are non-conserved inserts that have not been shown. The conserved GTEx+ motif of the scd6p family is shaded red. The residues involved in RNA binding are denoted by '#'s on the top of the alignment. The conserved C-terminal extension of the Scd6p family is shown in a box. The sequences are denoted by their gene name followed by the species abbreviation and GenBank Identifier (gi). The species abbreviations are: Af – *Archaeoglobus fulgidus*; Ec – *Escherichia coli*; Sau – *Staphylococcus aureus*; Afum – *Aspergillus fumigatus*; At – *Arabidopsis thaliana*; Cbr – *Caenorhabditis briggsae*; Ce – *Caenorhabditis elegans*; Dm – *Drosophila melanogaster*; Hs – *Homo sapiens*; Nc – *Neurospora crassa*; Pf-*Plasmodium falciparum*; Pwal – *Pleurodeles waltl*; Sc – *Saccharomyces cerevisiae*; and Sp – *Schizosaccharomyces pombe*.

packs against the uracil in the RNA, while backbones of the subsequent residues make hydrogen bonds with the base as well as the backbone of the RNA [25]. The conservation of this positively charged residue in the Scd6p family suggests that it may interact with the bases in RNA similar to the canonical archaeal and eukaryotic Sm domains [30] (Fig. 1). In the N-terminal half, the canonical Sm domains contain a conserved asparagine that makes a hydrogen bonding interaction with the uracil in the target RNA. This asparagine is typically replaced by a highly conserved threonine in the Scd6p family [25]. While the hydroxyl group of this residue might form a hydrogen bond with the base, it is unclear if it could confer the uracil-specificity that is provided by the asparagine in the canonical Sm domains. The Scd6p family has a polar residue instead of the aromatic residue that stacks against the base in most other canonical Sm domains (H37 in Af Sm1/AF0875; Fig. 1). This polar residue is likely to form hydrogen bonds with base rather than the stacking interactions which are observed in most other Sm domains [25]. These differences, along with the Scd6p family-specific GTEx+ motif that occurs between the N-

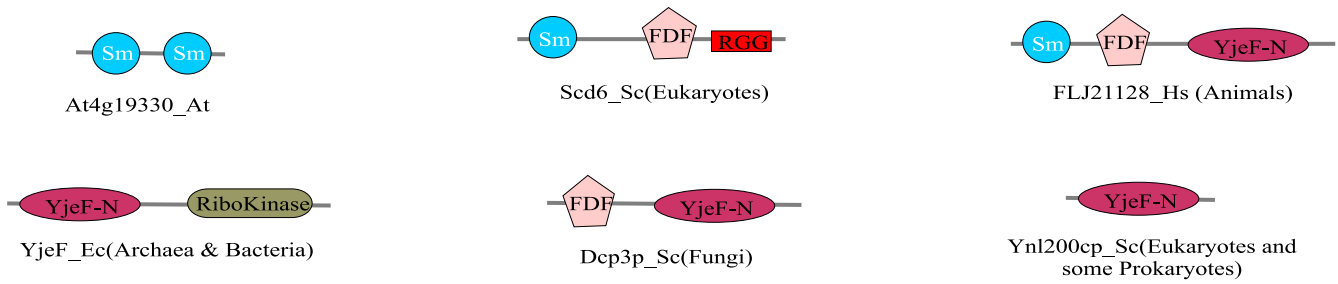

Figure 2
Domain architectures of Scd6p and FDF domain proteins. The domain architectures of the proteins containing the Scd6p, FDF and Yjef-N domains are shown. The representative protein name, organism and the phyletic pattern are given below the protein. The globular domains are drawn approximately to scale.

and C-terminal conserved regions of the domain, are likely to confer certain unique nucleic-acid-binding properties on the Scd6 family [30].

Most members of the Scd6p family contain a single Sm-related N-terminal domain fused to another conserved C-terminal domain, except At4g19330 from *Arabidopsis*, which is comprised of just two tandem repeats of the Sm domain (Fig. 2). In order to investigate the distinct C-terminal domain of the Scd6p family we initiated PSI-BLAST searches with this domain. In addition to members of the Scd6p family, these searches also recovered other proteins with significant e-values such as the Dcp3p (Yel015wp) protein from *S. cerevisiae* and its fungal relatives and uncharacterized proteins such as FLJ21128 (gi: 19923613) from *Homo sapiens* and its relatives from various animal clades. For example, searches with C-terminal domain of the human Scd6p ortholog (gi: 13559033) recovered Yel015wp/Dcp3 in iteration 5 with e = .003 and FLJ21128; e = $4*10^{-4}$. Reciprocal searches with this region from the above-mentioned proteins, such as Dcp3p and FLJ21128 recovered *bona fide* members of the Scd6p family with significant e-values (e.g. the region from FLJ21128 recovered the Rap55 in iteration 3; e = $2*10^{-4}$). Unlike the Scd6p family, this conserved region occurred in the N-terminal region of the Dcp3p and FLJ21128 proteins. These latter proteins additionally contained a C-terminal globular domain, which belongs of a specialized family of Rossmann fold domains. This family of Rossman fold domains also includes the N-terminal domain of the *E. coli* YjeF protein and, hereinafter we refer to this domain as the YjeF-N type Rossmann fold domains (see below for further discussion).

The above observations indicated that the conserved region shared by the Scd6p family, Yel015wp/Dcp3p and FLJ21128 is likely to define a novel domain. We named it the FDF domain after the characteristic signature that is present at N-termini of these domains (Fig. 3). The multiple alignment of the FDF domain shows that it is enriched in polar and charged residues with few hydrophobic residues embedded in their midst. It is predicted to adopt an entirely α-helical structure with multiple exposed hydrophilic loops. These features suggest that the FDF domain is likely to interact with RNA or highly charged peptides that are commonly found in the ribonucleoprotein complexes. Though the animal FLJ21128-like proteins and the fungal Yel015wp/Dcp3p differ in their architectures and are considerably divergent in terms of sequence, the presence of a shared architectural core (FDF domain fused to a YjeF-N-like Rossmann fold domain), which is not found in any other eukaryotic proteins suggests that they might belong to the same orthologous lineage shared by animals and fungi (Fig. 2 and 3).

N-terminal to the FDF domain, the FLJ21128-like proteins from animals, but not the fungal Dcp3p-like proteins, contain an additional small conserved globular domain. Based on its predicted secondary structure it is likely to adopt an all β-fold. Further analysis of this globular domain using profiles for conserved domains showed that it gave a significant hit (e-value=.005–001) with the Sm domain profile. This observation, taken together with its conservation pattern suggests that the extreme N-terminal domain in the FLJ21128-like proteins is yet another uncharacterized, divergent version of the Sm fold (Fig. 1 and 2).

Potential functions for the FDF and Scd6p-like Sm domain proteins in cell-cycle regulation and decapping

Genetic studies on *S. cerevisiae* Sdc6p and *S. pombe* Sum2p have been fairly opaque with regards to their functions. The Scd6p has been recovered as a suppressor of clathrin deficiency [31]. However, there is no evidence that it directly functions in the assembly of clathrin-coated vesicle. High-throughput localization studies have indicated

Figure 3

A multiple alignment of the FDF domain. Multiple sequence alignment of the FDF domain was constructed as described in Figure 1. In the secondary structure H represents a helix. The species abbreviations are as given in Figure 1 and additionally Ani – *Aspergillus nidulans*; Gze – *Gibberella zeae*; Mgr – *Magnaporthe grisea*.

that it is localized to the cytoplasm and not the nucleus in *S. cerevisiae* [32]. Sum2p was recovered as a weak suppressor of the over-production of the G2/M checkpoint regulator, Cdc25p [31]. The Cdc25p phosphatase is an activator of the cyclin dependent kinase Cdk2p and when over-produced it results in a bypass of the G2/M checkpoint, which ensures that DNA replication is completed before the M phase is initiated. Specifically, expression of the N-terminal Sm-like domain of Sum2p, but not the full length Sum2p, was found to restore the G2/M checkpoint bypass in Cdc25p-overproducing cells, as well as in cells with mutations in Cdk2p and Wee1p, which show identical checkpoint defects [31]. Consistent with these observations, the abrogation of the expression of the *C. elegans* ortholog of Sum2p, Y18D10A.17, results in cytokinesis defects and loss of fertility [33,34]. In cluster-analysis of gene expression patterns in *C. elegans*, Y18D10A.17 strongly groups with several genes that are over-expressed in the germline, oocytes and during cell division [35]. The newt homolog of Scd6p and Sum2p, Rap55 has been shown to be localized to mRNA containing cytoplasmic RNP particles [36]. It is present in a sharp temporal window in the oocytes, eggs and very early cleavage stages but not in the later stages of embryonic development or the adult tissues [36]. These observations point to a possible general role for these proteins in the regulation of pathways associated with cell-cycle progression.

The previously characterized Sm domain proteins in yeast have been shown to form at least three major hetero-heptameric complexes [35,37,38]. The first of these is a complex formed by the classical Sm proteins B, D1, D2, D3, E, F, and G and constitutes the core of the RNPs that bind the U1, U2, U4 and U5 snRNAs. A second complex formed by proteins Lsm2-8p is associated with the U6 snRNA and is a component of the U4/U6 and U4/U6 · U5 snRNPs. The third complex, consisting of Lsm1-7p, is associated with proteins like Dcp1p, Pat1p and Xrn1p, and is involved in RNA degradation via the decapping pathway [35,39,40]. Another heptameric nuclear Sm complex probably identical to the classical Sm complex of the spliceosomal complex is associated with the telomerase RNA subunit and is required for the telomerase function [41]. The conserved cytoplasmic localization of the Scd6p family and the association of Rap55 with mRNA containing particles resembles that of the processing bodies that contain the Lsm1-7p complex. This strongly suggests that the Scd6p proteins function in the cytoplasm, possibly as an alternative monomeric unit in formation of specialized Lsm1-7p-like heptameric complexes. These Scd6p-containing complexes could potentially bind a distinct subset of mRNAs that are specifically recognized by the Scd6p Sm-like domain. These Scd6p-containing complexes could possibly either target bound mRNAs for degradation or, conversely, stabilize the mRNAs by blocking their association with the Lsm1-7p complex involved in decapping. Under such a scenario, the specific regulation of the stabilities of various mRNAs encoding proteins involved in cytokinesis, cell cycle check points or clathrin coated vesicle assembly could account for the defects observed in these pathways. Interestingly, in line with this proposal, a second stronger suppressor of the checkpoint bypass caused by the over-

production of Cdc25p in *S. pombe* is the Sum3 gene, which encodes a RNA helicase [31]. Hence, it is possible that Sum2p and Sum3p act together to regulate the stability and translation of a similar set of mRNAs encoding check point proteins.

The available evidence also implicates the Dcp3p and FLJ21128 proteins with FDF and YjeF-N-type Rossmann fold domains in the decapping process. High-throughput analyses of protein-protein interactions in yeast using affinity precipitation and two-hybrid systems have consistently recovered the decapping enzymes Dcp1 and Dcp2, Dhh1p, the superfamily II helicase involved in decapping process, and the ribosomal protein S28 as potential interaction partners of Dcp3p [42-44]. The subcellular localization pattern of Dcp3p based on GFP tag analysis indicates that it is entirely cytoplasmic like Scd6p, Dhh1p. Specifically, it translocates to punctate foci [32], just like the decapping enzymes Dcp1p and Dcp2p and the Lsm1-7p complex [40,45]. These observations suggest that the Dcp3p and FLJ21128 proteins are likely to be associated with other proteins of the mRNA decapping complex in the specialized cytoplasmic processing bodies [45]. The presence of the N-terminal Sm domain in the FLJ21128 (and it orthologs from other animals) suggests that it might directly interact with other Sm proteins to be incorporated in specialized Sm heptamers.

Further clues regarding the functions of the Dcp3p and FLJ21128 are furnished by an analysis of the C-terminal YjeF-N-type Rossmann fold domain. Both iterative sequence searches with the PSI-BLAST program and structural similarity searches of PDB show that the dehydrogenase-type Rossmann domains are their closest relatives. For example a PSI-BLAST search with the YjeF-N domain of Dcp3p recovers dehydrogenases with significant e-values ($e = 10^{-5}$; iteration 6), while Ynl200cp (PDB:1jzt), a member of this family, recovers oxidoreductases like D-glycerate dehydrogenase with significant Z-scores ($Z = 8.9$) in structural similarity searches with the DALI program. However, a comparison of the sequence conservation pattern of the YjeF-N domains with that of the conventional Rossmann-fold dehydrogenases reveals several notable differences (Fig. 4 and Additional file 1). These include: 1) All members of this family contain two additional consecutive N-terminal helices that precede the first strand of the α/β core of the Rossmann fold and the core itself contains eight α/β units. Both these helices contain nearly absolutely conserved acidic residues. 2) The α/β core contains two characteristic aspartates; an absolutely conserved D at the end of strand 5 and one nearly universal D at end of strand 4. 3) The first helix of the α/β core of the Rossmann fold is extended by a whole turn resulting in the abbreviation of the glycine-rich nucleotide binding loop of the fold (Fig. 4). 4) The central sheet of the Rossmann fold is highly curved to form a peculiar barrel-like structure and the second additional N-terminal helix and the first helix of the α/β core pack against each other (Fig. 4). This structural quirk is chiefly stabilized by two sets of highly conserved interactions. Firstly, the salt-bridge and hydrogen-bonding interaction between the conserved acidic residue in the second N-terminal additional helix and the RH doublet in the first helix of the α/β core helps to positioning these two helices against one side of the curved sheet. Secondly, the hydrogen bonding between the conserved asparate at the end of strand 4 and the nearly absolutely conserved threonine C-terminal to strand 5 help in stabilizing the curvature of the central sheet (Fig. 4). 5) The acidic residue in the N-terminal-most additional helix of the YjeF-N, the acidic residue at the end of strand 5 and the polar residue (usually asparagine) from loop between strand 1 and helix 1 of the α/β core, line the mouth of the barrel- like structure to constitute the potential active site of this domain (Fig. 4).

In bacteria the YjeF-N domain is often found fused to a C-terminal kinase domain of the ribokinase superfamily (Fig. 2). Given that kinase domains are often fused to different phosphoesterase (phosphatase) domains [46], it is possible that the YjeF-N-type Rossmann fold domains may also catalyze this reaction. The conservation of the acidic residues in the predicted active site of the YjeF-N domains is reminiscent of the presence of such residues in the active sites of diverse hydrolases. Thus, in the context of the decapping pathway, it is possible that the YjeF-N domains of Dcp3p and FLJ21128 catalyze hydrolytic RNA-processing reactions, such as, phosphoester hydrolysis, dephosphorylation, demethylation or glycosyl bond hydrolysis.

The crystal structures of the archaeal Sm protein, SmAP3, and MscS provide examples of Sm domain toroids with additional N-terminal and/or C-terminal domains [29,47]. These structures indicate that these extension project out on either side of the of the central heptameric toroid formed by Sm domains [29,47]. If the Scd6p were to form similar toroidal structures, then the N- and C-terminal charged extensions with RG motifs and the FDF domains of the proteins are likely to project out similarly. In the canonical Sm toroids the RNA is threaded through the central cavity of this toroid, and previous studies have suggested that the charged extensions projecting away from the Sm core may form additional non-specific contacts with the RNA [25,26,48]. A similar RNA-binding function can be envisaged for the FDF domain. However, it is also possible that it forms a distinct interaction surface to bind charged peptides from proteins belonging to a specific RNP complex, possibly the complex that is involved in decapping [45].

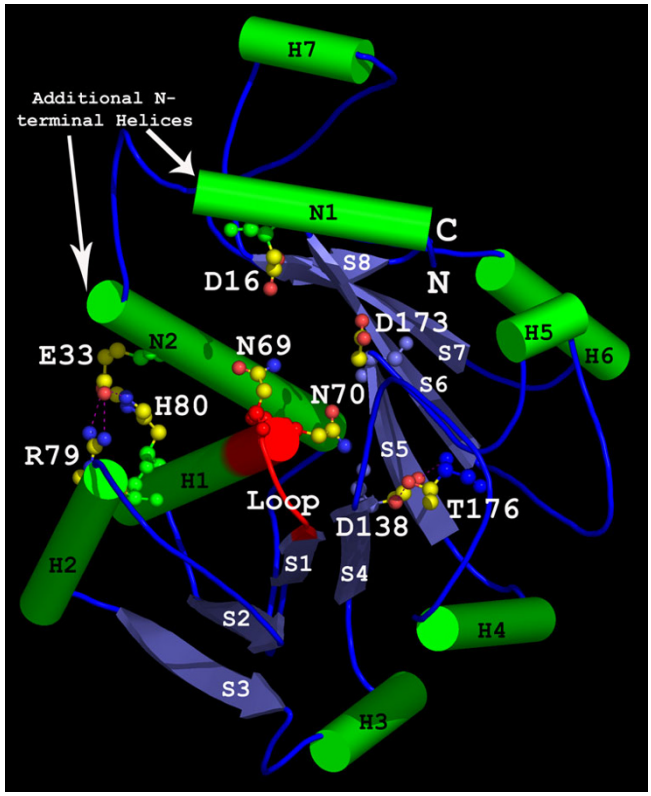

Figure 4

A Cartoon representation of the YjeF-N type Rossmann fold and its conserved features. The cartoon representation of the YjeF-N-type Rossmann fold domain was constructed using the crystal structure of the yeast YjeF-N domain containing protein (PDB: 1JZT). The N terminal helices are named N1 and N2, and the core helices and strands are named H1 to H7 and S1 to S8 respectively. The conserved residues of this fold corresponding to D16, E33, N69, N70, R79, H80, D138, D173 and T176 in this fold are shown in ball and stick representation. The salt bridges (E33 and R79 and H80) and hydrogen bonds (D138 and T176) between these conserved residues that are critical for the stabilization of the fold are shown as magenta dotted lines. The region between the strand 1 and helix 1 of the α/β core that corresponds to the glycine-rich nucleotide binding loop in the classic Rossmann fold (residues 66 and 72) is shown in red. Note the curvature of the central sheet and the packing of helix 1 of the α/β core and the second N-terminal additional helix.

Scd6p and Dcp3p in the context of the origin and evolution of the decapping machinery

The provenance of the decapping-dependent RNA degradation system in eukaryotes appears to have involved a number of different innovations and recruitment events. One process involved the *de novo* "invention" of new α-helical domains that mediate particular interactions, which are specific to this system. The most prominent of these inventions are the FDF domain and PATADs (for PAT1 alpha helical domains), the conserved α-helical domains seen in yeast Pat1p and its relatives from other eukaryotes. Sequence analysis and structure prediction also suggests that the decapping proteins, Edc1p/Edc2p [53], are also potential examples of poorly structured proteins that appear to be *de novo* innovations of the eukaryotes. In other instances, distinctive variants of preexisting globular folds appear to have been recruited for novel functions. An example of this is the decapping enzyme subunit Dcp1p, which contains a divergent variant of the peptide-binding EVH1 domain [54] that appears to have been recruited for a different, possibly catalytic function in the decapping process.

The MutT domain of Dcp2p [55] and the YjeF-N domain of Dcp3p appear to represent cases where the ancestral active site residues of the pre-existing catalytic domains appear to have been maintained, but they acquired a new set of substrates, specific to the decapping process. Analysis of phyletic patterns shows that Dcp2p is conserved throughout currently-sampled eukaryotes suggesting that it was present in the common ancestor of the extant eukaryotes. The closest relatives of this MutT domain are seen in bacteria, suggesting that the precursor of the Dcp2p catalytic domain may have been acquired very early in eukaryotic evolution via a transfer from a bacterial lineage. The precursor of Dcp3p and FLJ21128 was probably present at least since the common ancestor of the fungi and animals. Analysis of phyletic patterns of YjeF-N domains indicates that a second version of this domain, which is not fused to the FDF domain, is conserved across the three principal superkingdoms of life. Phylogenetic analysis of this version supports the monophyly of the YjeF-N domain in each of the three superkingdoms (barring certain lateral transfer involving bacteria; data not shown), suggesting that a single copy of the YjeF-N domain is traceable to the last universal common ancestor of all life forms. Its fusion to a small-molecule kinase of the ribokinase superfamily in bacteria suggests that the ancestral form of the YjeF-N domain may have functioned in the metabolism of a critical low molecular weight compound. The version of the YjeF-N domain found in Dcp3 and FLJ21128 was probably derived in the common ancestor of the animals and the fungi through duplication of the more ancient version of the YjeF-N domain. Alternatively, it could have been acquired via lateral transfer from a bacterial lineage. The extensive sequence divergence of the two versions currently prevents us from distinguishing between these possibilities through phylogenetic analysis.

The Sm domain is an ancient RNA binding domain that appears to have bound RNA ligands even in the last uni-

versal common ancestor of all extant life forms [1,24,37,49]. In bacteria, at least two ancient versions are present, namely Hfq [50,51] and the YhbC [52] (an uncharacterized protein found in most bacteria in the same operon with genes for the translation elongation factor NusA and initiation factor IF2; VA and LA, unpublished observations). Both these versions of the Sm domain are predicted to participate in binding RNAs in the context of translation. In archaea too the Sm domains interact with various RNA ligands, such as the RNAse P ribozyme [49].

The Sm superfamily of domains appears to have been vertically inherited by the eukaryotes from the common ancestor of the archaeo-eukaryotic lineage [1,37,49]. In eukaryotes the superfamily underwent a proliferation and appears to have been recruited as the core protein component of various eukaryote-specific RNP complexes such as the spliceosomal particles, the decapping complex and the telomerase complex. Phyletic patterns suggest that their explosive diversification in eukaryotes, giving rise to highly divergent forms such as the Scd6p family, appears to have happened prior to the divergence of the extant eukaryotic lineages. This suggests that the diversification of Sm-domain superfamily might have enabled them to interact with a diverse range of RNA ligands and protein partners and there by favored the emergence of multiple eukaryote-specific RNP complexes. Subsequently each of these complexes may have developed further, through the process of innovation of new α-helical domains and recruitment of catalytic domains from various sources.

Conclusions

We show that the Scd6p family contains a novel divergent version of the RNA-binding Sm domain and a previously uncharacterized C-terminal domain, the FDF domain. While the Scd6p Sm domain is predicted to bind RNA like most other prokaryotic and eukaryotic Sm domains, it is likely to have certain unique characteristics in terms of target specificity. The FDF domain is also present in several proteins such as Dcp3p and FLJ21128, where it is combined with the YjeF-N domain, a novel version of the Rossmann fold domain, and in some cases with another divergent version of the Sm domain. Along with other atypical Sm domains, like Ataxin-2 [24], Scd6 might form alternative Sm complexes, distinct from the, classical Sm, Lsm1-7p and Lsm2-8p complexes. A variety of contextual connections from expression, protein-protein interaction and intracellular localization data, suggest that the Scd6p, Dcp3p and FLJ21128 are associated with mRNAs in the cytoplasmic substructures and possibly regulate the stability of specific messages via the decapping system. The FDF domain may mediate interactions that are specific to these RNP complexes. Phyletic analysis of other components of the decapping system suggests that they have diverse origins and the explosive diversification of the Sm domains at the base of the eukaryotic radiation may have played an important role in the provenance of the uniquely eukaryotic RNP complexes.

Methods

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the BLASTP program [56]. Iterative database searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with the PSSM inclusion expectation (E) value threshold of 0.01 (unless specified otherwise); the searches were iterated until convergence [56,57]. For all searches with compositionally biased proteins, the statistical correction for this bias was employed. Multiple alignments were constructed using the T_Coffee [58] or PCMA [59] programs, followed by manual correction based on the PSI-BLAST results. Globular domains were predicted using the SEG program with the following parameters: window size 40, trigger complexity = 3.4; extension complexity = 3.75 [22]. All large-scale sequence analysis procedures were carried out using the SEALS package [60]. Specifically, pattern searches were carried out using the GREF program from this package. Structural similarity searches were conducted using the DALI program. The Swiss-PDB viewer [61] and Pymol programs were used to carry out manipulations of PDB files. Figures were rendered using PyMOL [62,63] or POV-Ray [64]. Protein secondary structure was predicted using a multiple alignment as the input for the PHD program [65,66]. Similarity-based clustering of proteins was carried out using the BLASTCLUST program [67].

Phylogenetic analysis was carried out using the maximum-likelihood methods. Maximum-likelihood distance matrices were constructed with the TreePuzzle 5 program [68] using 1000 replicates generated from the input alignment and used as the input for construction of neighbor-joining trees with the Weighbor program [69]. Weighbor uses a weighted NJ tree construction procedure that has been shown to effectively correct for long-branch effects [69]. Alternatively a full ML tree was constructed using the Proml program of the Phylip package [70]. This tree was used as the input tree to generate further full ML trees using the PhyML program [71] with 100 bootstrap replicates generated from the input alignment. The consensus of these trees was derived using the Consense program of the Phylip package to obtain the bootstrapped ML tree. Gene neighborhoods were determined by searching the NCBI PTT tables with a custom-written script. These tables can be accessed from the genomes division of the Entrez retrieval system [72].

Authors' contributions

VA contributed to the discovery process and preparation of the figures. LA conceived the study and contributed to the discovery process and preparation of the manuscript. Both authors read and approved the final manuscript.

Additional material

Additional File 1

Multiple alignment of the YjeF-N domain is provided in the supplementary material in the form of Additional file 1.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-5-45-S1.txt]

References

1. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002, **30:**1427-1464.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of Drosophila melanogaster.** *Science* 2000, **287:**2185-2195.
4. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV: **Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9:**608-628.
5. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci U S A* 1999, **96:**3801-3806.
6. Cerutti L, Mian N, Bateman A: **Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain.** *Trends Biochem Sci* 2000, **25:**481-482.
7. Spikes DA, Kramer J, Bingham PM, Van Doren K: **SWAP pre-mRNA splicing regulators are a novel, ancient protein family sharing a highly conserved sequence motif with the prp21 family of constitutive splicing proteins.** *Nucleic Acids Res* 1994, **22:**4510-4519.
8. Szymczyna BR, Bowman J, McCracken S, Pineda-Lucena A, Lu Y, Cox B, Lambermon M, Graveley BR, Arrowsmith CH, Blencowe BJ: **Structure and function of the PWI motif: a novel nucleic acid-binding domain that facilitates pre-mRNA processing.** *Genes Dev* 2003, **17:**461-475.
9. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.** *Genome Res* 2001, **11:**240-252.
10. Hall TA, Brown JW: **Archaeal RNase P has multiple protein subunits homologous to eukaryotic nuclear RNase P proteins.** *Rna* 2002, **8:**296-306.
11. Aravind L, Koonin EV: **THUMP - a predicted RNA-binding domain shared by 4-thiouridine and pseudouridine synthases and RNA methylases.** *Trends in Biochem Sci* 2001, **26:**215-217.
12. Clissold PM, Ponting CP: **PIN domains in nonsense-mediated mRNA decay and RNAi.** *Curr Biol* 2000, **10:**R888-90.
13. Blencowe BJ, Ouzounis CA: **The PWI motif: a new protein domain in splicing factors.** *Trends Biochem Sci* 1999, **24:**179-180.
14. Anantharaman Vivek., Koonin EV, Aravind L: **SPOUT: a class of methyltransferases that includes spoU and trmD RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases.** *J Mol Micro Biotech* 2002, **4:**71-75.
15. Aravind L, Koonin EV: **Novel predicted RNA-binding domains associated with the translation machinery.** *J Mol Evol* 1999, **48:**291-302.
16. Li H, Bingham PM: **Arginine/serine-rich domains of the su(wa) and tra RNA processing regulators target proteins to a subnuclear compartment implicated in splicing.** *Cell* 1991, **67:**335-342.

17. Birney E, Kumar S, Krainer AR: **Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors.** *Nucleic Acids Res* 1993, **21:**5803-5816.
18. Kiledjian M, Dreyfuss G: **Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box.** *Embo J* 1992, **11:**2655-2664.
19. Ramos A, Hollingworth D, Pastore A: **G-quartet-dependent recognition between the FMRP RGG box and RNA.** *Rna* 2003, **9:**1198-1207.
20. Anantharaman V, Koonin EV, Aravind L: **TRAM, a predicted RNA-binding domain, common to tRNA uracil methylation and adenine thiolation enzymes.** *FEMS Microbiol Lett* 2001, **197:**215-221.
21. Anantharaman V, Koonin EV, Aravind L: **SPOUT: a class of methyltransferases that includes spoU and trmD RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases.** *J Mol Microbiol Biotechnol* 2002, **4:**71-75.
22. Wootton JC: **Non-globular domains in protein sequences: automated segmentation using complexity measures.** *Comput Chem* 1994, **18:**269-285.
23. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
24. Neuwald AF, Koonin EV: **Ataxin-2, global regulators of bacterial gene expression, and spliceosomal snRNP proteins share a conserved domain.** *J Mol Med* 1998, **76:**3-5.
25. Toro I, Thore S, Mayer C, Basquin J, Seraphin B, Suck D: **RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex.** *Embo J* 2001, **20:**2293-2303.
26. Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K: **Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs.** *Cell* 1999, **96:**375-387.
27. B: **SCOP database.** [http://scop.mrc-lmb.cam.ac.uk/scop/].
28. Thore S, Mayer C, Sauter C, Weeks S, Suck D: **Crystal structures of the Pyrococcus abyssi Sm core and its complex with RNA. Common features of RNA binding in archaea and eukarya.** *J Biol Chem* 2003, **278:**1239-1247.
29. Bass RB, Strop P, Barclay M, Rees DC: **Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel.** *Science* 2002, **298:**1582-1587.
30. Achsel T, Stark H, Luhrmann R: **The Sm domain is an ancient RNA-binding motif with oligo(U) specificity.** *Proc Natl Acad Sci U S A* 2001, **98:**3685-9. Epub 2001 Mar 20..
31. Forbes KC, Humphrey T, Enoch T: **Suppressors of cdc25p overexpression identify two pathways that influence the G2/M checkpoint in fission yeast.** *Genetics* 1998, **150:**1361-1375.
32. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
33. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J: **Functional genomic analysis of C. elegans chromosome I by systematic RNA interference.** *Nature* 2000, **408:**325-330.
34. Simmer F, Moorman C, Van Der Linden AM, Kuijk E, Van Den Berghe PV, Kamath R, Fraser AG, Ahringer J, Plasterk RH: **Genome-Wide RNAi of C. elegans Using the Hypersensitive rrf-3 Strain Reveals Novel Gene Functions.** *PLoS Biol* 2003, **1:**E12. Epub 2003 Oct 13..
35. Bouveret E, Rigaut G, Shevchenko A, Wilm M, Seraphin B: **A Sm-like protein complex that participates in mRNA degradation.** *Embo J* 2000, **19:**1661-1671.
36. Lieb B, Carl M, Hock R, Gebauer D, Scheer U: **Identification of a novel mRNA-associated protein in oocytes of Pleurodeles waltl and Xenopus laevis.** *Exp Cell Res* 1998, **245:**272-281.
37. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Seraphin B: **Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin.** *Embo J* 1999, **18:**3451-3462.
38. Raker VA, Plessel G, Luhrmann R: **The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro.** *Embo J* 1996, **15:**2256-2269.
39. Ingelfinger D, Arndt-Jovin DJ, Luhrmann R, Achsel T: **The human LSm1-7 proteins colocalize with the mRNA-degrading enzymes Dcp1/2 and Xrnl in distinct cytoplasmic foci.** *Rna* 2002, **8:**1489-1501.
40. Tharun S, He W, Mayes AE, Lennertz P, Beggs JD, Parker R: **Yeast Sm-like proteins function in mRNA decapping and decay.** *Nature* 2000, **404:**515-518.
41. Seto AG, Zaug AJ, Sobel SG, Wolin SL, Cech TR: **Saccharomyces cerevisiae telomerase is an Sm small nuclear ribonucleoprotein particle.** *Nature* 1999, **401:**177-180.
42. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403:**623-627.
43. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci U S A* 2001, **98:**4569-74. Epub 2001 Mar 13..
44. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
45. Sheth U, Parker R: **Decapping and decay of messenger RNA occur in cytoplasmic processing bodies.** *Science* 2003, **300:**805-808.
46. Leipe DD, Koonin EV, Aravind L: **Evolution and classification of P-loop kinases and related proteins.** *J Mol Biol* 2003, **333:**781-815.
47. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D: **Structure and assembly of an augmented Sm-like archaeal protein 14-mer.** *Proc Natl Acad Sci U S A* 2003, **100:**4539-44. Epub 2003 Mar 31..
48. Zhang D, Abovich N, Rosbash M: **A biochemical function for the Sm complex.** *Mol Cell* 2001, **7:**319-329.
49. Schwartz D, Decker CJ, Parker R: **The enhancer of decapping proteins, Edc1p and Edc2p, bind RNA and stimulate the activity of the decapping enzyme.** *Rna* 2003, **9:**239-251.
50. Callebaut I: **An EVH1/WH1 domain as a key actor in TGFbeta signalling.** *FEBS Lett* 2002, **519:**178-180.
51. Dunckley T, Parker R: **The DCP2 protein is required for mRNA decapping in Saccharomyces cerevisiae and contains a functional MutT motif.** *Embo J* 1999, **18:**5411-5422.
52. Toro I, Basquin J, Teo-Dreher H, Suck D: **Archaeal Sm proteins form heptameric and hexameric complexes: crystal structures of the Sm1 and Sm2 proteins from the hyperthermophile Archaeoglobus fulgidus.** *J Mol Biol* 2002, **320:**129-142.
53. Schumacher MA, Pearson RF, Moller T, Valentin-Hansen P, Brennan RG: **Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein.** *Embo J* 2002, **21:**3546-3556.
54. Sauter C, Basquin J, Suck D: **Sm-like proteins in Eubacteria: the crystal structure of the Hfq protein from Escherichia coli.** *Nucleic Acids Res* 2003, **31:**4091-4098.
55. Yu L, Gunasekera AH, Mack J, Olejniczak ET, Chovan LE, Ruan X, Towne DL, Lerner CG, Fesik SW: **Solution structure and function of a conserved protein SP14.3 encoded by an essential Streptococcus pneumoniae gene.** *J Mol Biol* 2001, **311:**593-604.
56. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
57. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287:**1023-1040.
58. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-217.
59. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence alignment based on profile consistency.** *Bioinformatics* 2003, **19:**427-428.
60. A: **SEALS package.** [http://www.ncbi.nlm.nih.gov/CBBresearch/Walker/SEALS/index.html].

61. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-Pdb-Viewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.
62. DeLano WL: **The PyMOL Molecular Graphics System.** San Carlos, CA, USA, DeLano Scientific; 2002.
63. A: **Pymol.** [http://www.pymol.org].
64. A: **PovRay.** [http://www.povray.org/].
65. Rost B, Fariselli P, Casadio R: **Topology prediction for helical transmembrane proteins at 86% accuracy.** *Protein Sci* 1996, **5:**1704-1718.
66. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584-599.
67. A: **BLASTCLUST.** [ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.txt].
68. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
69. Bruno WJ, Socci ND, Halpern AL: **Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction.** *Mol Biol Evol* 2000, **17:**189-197.
70. Felsenstein J: **PHYLIP -- Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
71. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696-704.
72. B: **Gene Neighborhood Tables.** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome].